Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky772

A mating-type mutagenesis screen identifies a zinc-finger protein required for specific DNA excision events in *Paramecium*

Simran Bhullar[1,2,*], Cyril Denby Wilkes[3], Olivier Arnaiz[3], Mariusz Nowacki[2], Linda Sperling[3] and Eric Meyer[1,*]

[1]IBENS, Ecole Normale Supérieure, CNRS, Inserm, PSL University, F-75005 Paris, France, [2]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland and [3]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette cedex, France

Received June 15, 2018; Revised August 11, 2018; Editorial Decision August 14, 2018; Accepted August 24, 2018

ABSTRACT

In the ciliate *Paramecium tetraurelia*, functional genes are reconstituted during development of the somatic macronucleus through the precise excision of ∼45 000 single-copy Internal Eliminated Sequences (IESs), thought to be the degenerate remnants of ancient transposon insertions. Like introns, IESs are marked only by a weak consensus at their ends. How such a diverse set of sequences is faithfully recognized and precisely excised remains unclear: specialized small RNAs have been implicated, but in their absence up to ∼60% of IESs are still correctly excised. To get further insight, we designed a mutagenesis screen based on the hypersensitivity of a specific excision event in the *mtA* gene, which determines mating types. Unlike most IES-containing genes, the active form of *mtA* is the unexcised one, allowing the recovery of hypomorphic alleles of essential IES recognition/excision factors. Such is the case of one mutation recovered in the Piwi gene *PTIWI09*, a key player in small RNA-mediated IES recognition. Another mutation identified a novel protein with a C2H2 zinc finger, mtGa, which is required for excision of a small subset of IESs characterized by enrichment in a 5-bp motif. The unexpected implication of a sequence-specific factor establishes a new paradigm for IES recognition and/or excision.

INTRODUCTION

Site-specific DNA recombination occurs on a massive scale during the development of ciliates, making them ideal models to study the genetic and epigenetic regulation of programmed genome rearrangements in eukaryotes. These unicellular organisms harbor two kinds of nuclei within the same cytoplasm. The diploid micronucleus (MIC) is transcriptionally silent and only serves germline functions, undergoing meiosis during sexual events, while the polyploid macronucleus (MAC) is a somatic nucleus which divides by a non-mitotic process during vegetative growth and is responsible for all gene expression, but is not passed on to sexual progeny. Following MIC meiosis and fertilization, the parental MAC is discarded and a new one develops from a mitotic copy of the diploid zygotic nucleus. MAC development involves endoreplication to reach the final ploidy level (∼800n in *Paramecium tetraurelia*) as well as extensive and reproducible rearrangements, including the elimination of transposable elements (TEs) and satellite DNA and the fragmentation of chromosomes (1).

In *P. tetraurelia*, the rearrangement program also includes the precise excision of ∼45 000 short, single-copy Internal Eliminated Sequences (IESs) dispersed throughout the germline genome (2). IESs are invariably bounded by two 5′-TA-3′ dinucleotides which precisely recombine into one upon excision. IESs are thought to be the degenerate remnants of ancient TE insertions that have been tolerated in the germline because of the evolution of a highly efficient and precise mechanism for somatic excision (3). Indeed, 82% of the IESs are inserted within protein-coding genes, and about half of the ∼40 000 genes are interrupted by IESs that must be precisely removed before functional proteins can be expressed. IES excision in the developing MAC depends on a domesticated piggyBac transposase, Pgm, which is required for the introduction of staggered double-strand breaks (DSBs) at both IES ends (4,5). The highly non-random size distribution of IESs shorter than 150 bp, which forms a series of peaks spaced by ∼10.2 bp, suggests structural constraints on excision (2). DSBs produce 4-nt 5′ overhangs centered on the TAs, and flanking se-

*To whom correspondence should be addressed. Tel: + 33 1 44 32 39 48; Fax: + 33 1 44 32 39 41; Email: emeyer@biologie.ens.fr
Correspondence may also be addressed to Simran Bhullar. Email: simrnbhullar@gmail.com

quences are then rejoined precisely on the overhanging TAs by the Non-Homologous End Joining pathway (6,7).

In contrast to the molecular details of DNA recombination, the mechanisms ensuring the faithful recognition, in each sexual generation, of such a large number of unique sequences remain poorly understood. A weak consensus adjacent to the TAs, 5′-(TA)YAGYNR-3′, is often present at both IES ends, in opposite orientations (2,8). Although it may guide the choice of specific TA boundaries for many IESs, the consensus is not a strict requisite for excision and does not contain sufficient information to explain the specificity of the IES excision pattern genome-wide.

A partial answer to the specificity problem was provided by the finding that scnRNAs, a meiosis-specific class of small RNAs, are required for correct rearrangements in the developing MAC. Initially produced from the entire germline genome during MIC meiosis by the Dicer-like proteins Dcl2 and Dcl3 (9–11), scnRNAs are thought to mediate a genomic subtraction that identifies IESs as being absent from the parental MAC genome, rearranged in the previous generation. After being loaded on the Piwi proteins Ptiwi01 and Ptiwi09 (12,13), scnRNAs are imported into the parental MAC and those able to pair with pervasive nascent transcripts are somehow inactivated (14). Thus, only those originating from MIC-specific sequences would be licensed to later target the same sequences in the developing zygotic MAC, by pairing with TFIIS4-dependent nascent transcripts (15). In both parental and zygotic MACs, pairing of scnRNAs with nascent transcripts is thought to involve the RNA-binding and Ago-hook-containing proteins Nowa1/2 (16,17). In the zygotic MAC, scnRNA pairing would trigger the methylation of histone H3 on K9 and K27 by the SET-domain protein Ezl1 (18), a step which also requires the chromatin assembly factor Pt-CAF1 (19). The resulting histone marks, rather than IES sequences, would then recruit the Pgm-containing excision complex to the correct sites.

The scnRNA pathway could in principle account for the recognition of all IESs. The model further explains how experimental introduction of some IESs into the parental MAC, by inactivating homologous germline scnRNAs at the following meiosis, can inhibit excision of the same sequences in sexual progeny (20,21). The same mechanism can also program the deletion of cellular genes that happen to be missing in the parental MAC (1), and has been naturally co-opted to mediate the transgenerational epigenetic inheritance of mating types (see below). However, genome-wide studies have shown that the scnRNA pathway is only modestly involved in the recognition of IESs. Indeed, while RNAi-mediated depletion of Dcl2 and Dcl3 (or Ptiwi01 and Ptiwi09) strongly impairs elimination of TEs, only a small fraction of all IESs (<10%) are significantly retained in the new MAC genome, and at quantitatively different levels (fractions of polyploid copies retaining the IES) (11,13,17,18).

A second class of small RNAs called iesRNAs, produced by the Dicer-like protein Dcl5, is required for complete excision of a similarly small (and partially overlapping) subset of IESs (11,13,22). iesRNAs are thought to be produced from IESs after their excision and to act in a positive feedback loop ensuring complete IES excision in the polyploid MAC. The simultaneous depletion of scnRNAs and iesRNAs impairs excision of a larger subset; up to ∼40% of IESs are affected to some extent, suggesting that iesRNAs can somehow compensate for the lack of scnRNAs, thus defining a 'small RNA-dependent' subset that is fairly consistent with the subsets of IESs retained in the knockdowns of TFIIS4 or Nowa1/2 (17). Depletion of Ezl1, however, affects excision of ∼70% of IESs, including the small RNA-dependent 40% and an additional 30% that appear to be targeted by Ezl1 independently of small RNAs (17,18). How this is achieved, and how the last 30% of IESs are recognized independently of small RNAs and Ezl1-mediated H3 modifications, is currently unknown.

A mutagenesis screen could provide new insight into these questions by identifying relevant factors without any preconceived idea as to their nature. *P. tetraurelia* is ideally suited for genetic analyses, but mutations affecting any step of IES excision are expected to be lethal. The only Mendelian mutant ever found to be affected in programmed genome rearrangements, mtF^{E}, was initially isolated as being constitutively determined for one of the two mating types (23,24). In the wild type, mating types are determined during MAC development by alternative rearrangements of the *mtA* gene, which encodes a transmembrane protein involved in cross-recognition of mating types E (Even) and O (Odd) (25). *mtA* is expressed only in mating type E clones because its promoter is excised as an IES during development of mating type O clones. In contrast to other IESs, the *mtA* promoter is a functional part of a cellular gene; the IES excision machinery has in this case been co-opted to inactivate the gene. The scnRNA pathway ensures that this occurs only when the *mtA* promoter is absent from the parental MAC, resulting in maternal (cytoplasmic) inheritance of mating types after conjugation or autogamy (a self-fertilization sexual process) (25).

Developmental excision of the *mtA* promoter in mating type O, like other cases of maternally inherited deletions of genes or gene parts, is more sensitive to defects of the scnRNA pathway than is the excision of gene-interrupting IESs. For instance, the single silencing of either *PTIWI01* or *PTIWI09*, which are redundant ohnologs from the last whole-genome duplication (WGD), is sufficient to cause retention of the *mtA* promoter in sexual progeny of mating type O cells, while both must be silenced simultaneously to observe detectable retention of scnRNA-dependent IESs (12,13,25). This observation inspired the design of a mutagenesis screen for factors involved in IES recognition and/or excision, based on the appearance of E cells in the sexual progeny of mutagenized O cells. We reasoned that hypomorphic mutations with only weak effects on the excision of most IESs might preserve cell viability while still allowing the expression of mating type E, a selectable phenotype. We recovered two mutations, both in genes with redundant ohnologs from the last WGD. One is a hypomorphic allele of *PTIWI09* and the other encodes a novel zinc-finger protein required for excision of a very small subset of IESs. The latter is the first example of a variant mechanistic requirement linked to the presence of a specific sequence motif in IESs.

MATERIALS AND METHODS

Paramecium strains and cultivation

Unless otherwise stated, all experiments were carried out with the entirely homozygous reference strain 51 of *P. tetraurelia*. Strain 138 of *P. octaurelia* was from the stock collection of the Centre of Core Facilities 'Collections of Microorganisms' at St Petersburg State University, Russia. Unless otherwise stated, cells were grown at 27°C in a wheat grass powder (WGP) (Pines International) infusion medium bacterized with *Klebsiella pneumoniae* and supplemented with 0.8 mg/ml β-sitosterol (Merck) before use (26,27).

UV mutagenesis and screening for mating-type revertants

UV mutagenesis was carried out as described (28). Briefly, batches of ∼200 000 cells (wild-type mating type O, ≥26 divisions of clonal age) were irradiated with UV (254 nm) at a dose of 650 J/m^2, previously shown to result in 20–25% lethality in post-autogamous progeny and in an average of ∼30 mutations in the MAC genomes of viable progeny. Irradiated cells were grown for two divisions in the dark, starved to induce autogamy, and let starve for five more days to facilitate elimination of fragments of the maternal MAC. To take post-autogamous cultures through a second round of mass autogamy, ∼50 000 post-autogamous cells were refed with 400 ml of rich (1× WGP) bacterized medium and kept at 18°C to allow two divisions/day, and the same dilution was repeated every day until previous dilutions showed ≥70% autogamous cells upon starvation. Mass post-autogamous cultures (after the first or second round of autogamy) were then stored at 14°C. To screen these cell populations for mating type E revertants, a fraction was refed with rich medium at 27°C so that the culture would begin to starve again the next day; as cells became sexually reactive, small batches were observed under the binocular in Petri dishes to identify clumps of agglutinated cells. Aggregates were isolated in 250 μl of light (0.2× WGP) bacterized medium and dissociated with a glass micro-pipette, and each individual cell was again isolated in the same amount of medium in glass depression slides and grown at 27°C. Upon starvation, clones were tested for mating types to identify the original E revertant in each aggregate (successful for ∼50% of aggregates), and the E clone was then taken through an additional round of autogamy before genetic analysis; O clones were not further considered.

Mating type tests

Testers were prepared as described previously (25). Approximately 1000 autogamous cells of known mating types were fed 4 ml of bacterized light medium and incubated at 27°C overnight. The next day, tubes were refed with 8 ml of light medium and again incubated overnight at 27°C. Sexual reactivity was checked the following day by mixing small aliquots of complementary mating types. Mass post-autogamous progenies to be tested were made reactive in the same way. Individual clones to be tested were grown in 250 μl of light medium until starved (2–3 days), then refed 1 volume and tested the next day.

Genetic analysis of mating type E revertants

E revertant lines were crossed to O cells carrying a homozygous genetic marker, the recessive trichocyst non-discharge mutation *nd7-1* (29). After the separation of exconjugants, genetic exchange could thus be confirmed in the F1 clones deriving from the O parents (identified by mating-type tests) by complementation of the *nd7-1* phenotype. These F1 clones were then taken through autogamy, and the appearance of E progeny was checked by one of two alternative procedures. A complete, detailed scheme was used for revertants isolated in the first irradiation experiment: 30 independent F2 homozygotes were isolated from each post-autogamous population, grown and tested for mating types. Unless about half of them turned out to be E, indicating a mutation in a Zygotically Expressed Gene (ZEG), the whole set of F2 clones from each revertant was then taken through another round of autogamy. Post-autogamous F3 populations (∼800 cells) were refed for 2–3 divisions and tested *en masse* for mating types to check for possible mutations in Maternally Expressed Genes (MEGs).

A simplified and faster procedure was used for revertants isolated in the second irradiation experiment. Instead of isolating individual F2 clones, the post-autogamous progeny from each F1 was refed *en masse* for a few divisions and tested for mating types. The segregation of a mutant allele in a ZEG would translate into selfing, i.e. intra-population conjugation due to the presence of both genotypes. When entire F2 populations remained O, they were further grown *en masse* at 27°C, by daily dilutions of about ∼500 cells, until they were old enough for the second autogamy. Mass mating type tests were then repeated with F3 populations, to check for selfing that would indicate the presence of both mutant and wild-type alleles of a MEG.

Identification of mutations by whole-genome sequencing

We used the strategy designed in (28): after crossing the mutant to the wild type, an F1 heterozygote is taken through autogamy, and F2 homozygotes with a mutant phenotype are pooled to sequence their MAC genome. With a sufficient coverage, the phenotype-causing mutation is easily identified as being present in 100% of reads, while irrelevant, unlinked mutations will appear only in 50% of reads on average. For *mtGa-1*, the sequenced pool included 19 E F2 homozygotes from the second backcross. The MAC genome of the corresponding pool of 19 F3 populations was also sequenced. For *PTIWI09-1*, we sequenced the MAC genome of a pool of 27 F3 clones from the 3rd backcross, each randomly picked from a different selfing F3 population. Technical details and statistics for these datasets are given in Supplementary Table S5. Reads were then mapped to the MAC genome to identify candidate mutations as described (28).

IES retention analysis

Illumina reads were trimmed with a minimum quality score of 20 (sickle v1.2) and adapters were removed using cutadapt (v1.8.1). Reads were then mapped using bowtie2 v2.3.3 (30) against MAC and MAC+IES *P. tetraurelia* reference genomes (2). IES retention was assessed using PartIES (31). Briefly, the number of reads matching the IES-

containing forms were counted over each TA boundary, as well as the number of reads matching the IES-excised junction. The boundary-specific retention score is defined as the fraction of unexcised reads among all reads; i.e. a score of 1 means that only unexcised reads were detected. Read counts were compared to those from a control dataset using a frequency comparison test, based on a binomial law of probability. Resulting *P*-values were adjusted for multiple testing using the Benjamini & Hochberg method (32). An IES is considered to be significantly retained if at least one of its two boundary retention scores has a *P*-value < 0.05. In this study we used seven different control datasets (Supplementary Tabel S5): an IES was considered significantly retained if at least one of its boundaries was significantly more retained in the tested sample than in each of the seven control datasets.

Silencing of *mtG* genes by RNAi

dsRNA was produced from a 214-bp divergent portion of the coding sequences of *mtGa* and *mtGb* (48–261 from the ATG, 63.6% identity) to avoid cross-silencing. PCR fragments were cloned in vector L4440 and *Escherichia coli* strain HT115 DE3 as described (33). The homologous sequences were used for silencing of *mtGa* and *mtGb* in *P. octaurelia*, strain 138.

DNA extraction and PCR

Large-scale DNA samples for whole-genome sequencing were extracted from purified MACs as described (2). Mass post-autogamous progenies from cultures silenced for the *mtG* genes were allowed to undergo 1–2 divisions after autogamy to decrease the fraction of maternal MAC DNA. Small-scale DNA samples were prepared from $\leq$1,000 cells using either NucleoSpin Tissue Kit (Macherey-Nagel) or GenElute Mammalian Genomic DNA MiniPrep Kit (Sigma). PCR amplifications were performed with Phusion polymerase (ThermoFisher). PCR products to be sequenced were purified by Wizard SV Gel and PCR clean-up kit (Promega).

GFP fusion constructs and microinjection

The mtGa-GFP fusion construct under the control of endogenous regulatory sequences (pmtGa-GFP) contained 551 bp of MAC sequences upstream of the *mtGa* translation start and 165 bp downstream of the translation stop, and the EGFP coding sequence was inserted at the 5′ end of the coding sequence. In the pmtGa-GFP-L2 construct, the *mtGa* promoter and terminator sequences were replaced by those of the *PGML2* gene (PTET.51.1.G0380073; 557 bp upstream of the translation start and 233 bp downstream of the translation stop). Linearized plasmids were microinjected into the MAC of vegetative cells as described (34).

H3K9me3 immunofluorescence and confocal analysis

The immunocytochemistry protocol was adapted from (19,35). Cells were permeabilized for 10 min in 1× PHEM buffer (300 mM PIPES, 25 mM HEPES, 10 mM EGTA, 2 mM $MgCl_2$, pH 6.9) with 1% Triton-X-100, followed by fixation in 2% paraformaldehyde in 1× PHEM buffer for 10 min. Cells were then blocked in TBSTEM (10 mM Tris, 0.15 M NaCl, 10 mM EGTA, 2 mM $MgCl_2$, pH 7.4 with 1% v/v Tween 20) containing 3% BSA. Cells were either processed directly or stored at 4°C overnight before incubating with an anti-H3K9me3 rabbit polyclonal antibody (07-449, Millipore) diluted 1:200 in TBSTEM containing 3% BSA for 1 h, washed and incubated with Alexa Fluor 568-conjugated goat anti-rabbit IgG (A-11036, Invitrogen) at 1:200 for 1 h. Cells were stained with DAPI for 10 min and mounted in Citifuor. Images were acquired using a Leica Confocal SP5 laser-scanning microscope with 568, 488 and 405 laser line excitations for detection of Alexa Fluor 568 (red), GFP (green), and DAPI (blue), respectively. Z-series were performed with default Z-steps of 0.13 μm. Images were processed with the ICY software.

Bioinformatics web services used

DNA and protein sequences were aligned using either clustalW or MUSCLE at http://phylogeny.fr website (36). Detection of motifs in IES sequences used DREME at http://meme-suite.org/tools/dreme (37). The zinc-finger binding site prediction was carried out at http://zf.princeton.edu (38).

RESULTS

Isolation of mating type E revertants in the sexual progeny of mutagenized type O cells

Batches of ∼200 000 mating type O cells of strain 51 were irradiated with UV to induce heterozygous mutations in the MIC genome. Cells were then grown for two divisions and starved to trigger autogamy, which is expected to make one fourth of germline mutations homozygous in the new MICs and MACs of progeny. Indeed each vegetative cell contains two diploid MICs which both undergo meiosis, but only one of eight haploid nuclei is kept and undergoes an additional mitosis, yielding two genetically identical gametic nuclei. These fuse together during the self-fertilization step of autogamy, making the zygotic nucleus homozygous for one of the 4 parental MIC alleles at each locus. Two post-zygotic divisions then give rise to the new MICs and MACs.

One round of autogamy is sufficient to reveal the phenotypic effects of mutations in genes that are expressed only from the zygotic genome during development, such as *mtF* and a few others (13,24). A fraction of post-autogamous cell populations were therefore screened for E-expressing revertants by refeeding them enough for 2–3 divisions and observing them as they were beginning to starve and became sexually reactive: in that physiological state the rare E cells will agglutinate with up to 3–5 O cells, forming aggregates that cannot swim and sink to the bottom of the dish. Each aggregate was dissociated before the formation of conjugating pairs and individual cells were grown and tested for mating types to identify the E clone (see Materials and Methods).

Most of the genes known to play a role in programmed genome rearrangements, however, are expressed from the maternal genome rather than the zygotic one. Mutations in

such genes must be homozygous in the maternal MAC to affect excision of the *mtA* promoter in the zygotic MAC, and most of the screening was therefore done after a second round of autogamy. Because young clones cannot undergo autogamy before they are ∼20-division old, the post-autogamous progeny of mutagenized cells was cultured *en masse*, refeeding one fourth of the culture (∼50 000 cells) every two divisions, until starvation could trigger autogamy in ≥70% of cells. Second-generation homozygotes were then again grown for a few divisions and screened for mating-type revertants. In two experiments, a total of 162 revertants were isolated (of which only four after the first autogamy); from these, 60 stable E cell lines could finally be established.

Distinguishing Mendelian mutants from spontaneous mating-type revertants

An unknown fraction of these E revertants may be cases of spontaneous reversion, i.e. spontaneous retention of the *mtA* promoter in the developing MAC when its absence from the maternal MAC should normally license homologous scnRNAs to target its excision (25). The frequency of O-to-E epigenetic reversion at autogamy was previously estimated to be <1/50 000 (39), though this may vary among strains or even clones. To distinguish spontaneous revertants from those induced by germline mutations, the most direct approach is genetic analysis. When spontaneous E revertants are crossed with wild-type O cells, the newly established mating type E will be transmitted maternally to further sexual generations, i.e. only in the revertant's cytoplasmic line of descent (Figure 1A). In contrast, if mating type E was induced by a homozygous mutation, the reciprocal fertilization that occurs during conjugation will make both F1s heterozygous, and the two alleles will segregate 1:1 among F2 homozygotes, after autogamy of the F1s (Figure 1B). In the F1 deriving from the O parent, no change of mating type is expected for recessive mutations, but 50% of post-autogamous F2 clones will be mutant homozygotes that should switch to E if the mutation is in a Zygotically Expressed Gene (ZEG) (Figure 1C). If the mutation is in a Maternally Expressed Gene (MEG), all F2 clones should remain O, but half of them (the mutant homozygotes) should give rise to E F3 progeny (Figure 1D).

All 60 E revertant lines were therefore crossed to wild-type O cells, and the appearance of E clones in the F2 or F3 post-autogamous progeny from the mating type O F1s was tested by one of two alternative procedures (see Materials and Methods). In one case, about half of F2 clones developed as mating type E, suggesting a putative mutation in a ZEG. This was confirmed by two rounds of backcross to the wild type (Supplementary Table S1). However, ∼10% of O F2 clones gave rise to E progeny in the F3 generation, suggesting that some level of maternal expression from the heterozygous F1 MAC can occasionally prevent the switch to E in mutant F2s.

In one other case, all F2 clones were O but about half of them gave rise to F3 progenies that showed a 'selfer' phenotype (intra-population conjugation) when tested in mass for mating types, while F3 progenies from the other F2s remained O. This suggested a low-penetrance mutation in a MEG. Backcrosses confirmed the 1:1 Mendelian ratio of the two types of F2s, and the fact that only a fraction of individual F3 clones in the selfing progenies had switched to E (Supplementary Table S2). In the remaining 58 lines, the O F1 did not give rise to any E progeny in the F2 or F3 generations, suggesting that the rate of spontaneous O-to-

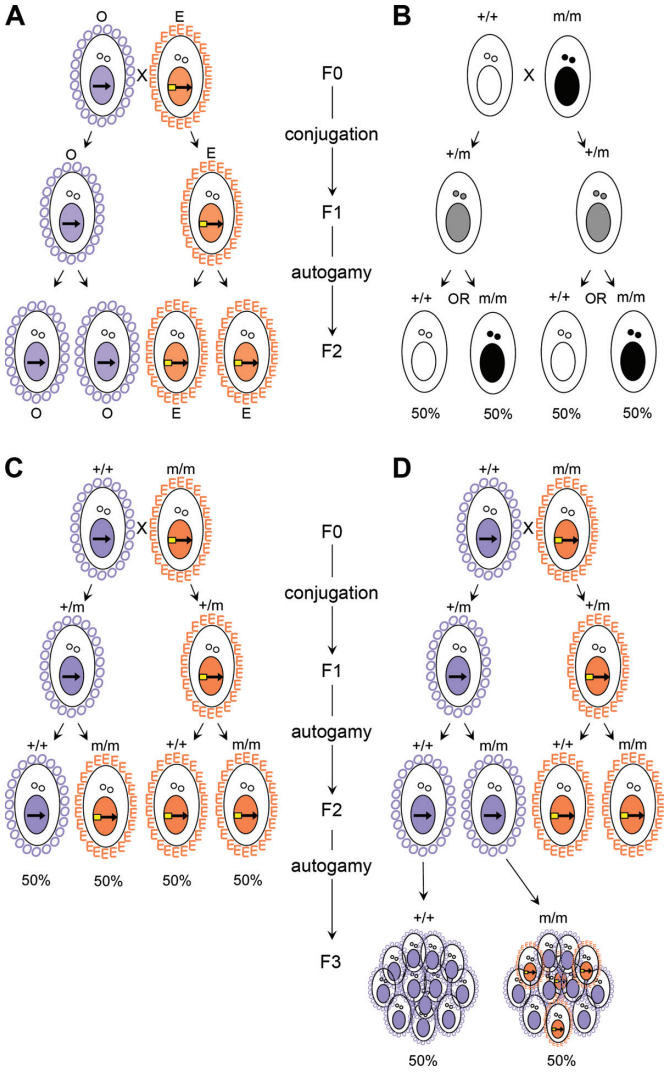

Figure 1. Genetic analysis of mating-type revertants. (**A**) Mating type E revertants resulting from spontaneous reversion behave as wild-type mating type E after a cross to O cells: mating type E is maternally transmitted at conjugation and autogamy, and the mating type O cytoplasmic lineage remains pure O. The arrow drawn in the MAC symbolizes the *mtA* gene, and the yellow box at its left end indicates clones that retain its promoter in the MAC genome, resulting in mating type E expression. (**B**) Mendelian segregation in a cross of wild-type (+) and mutant (m) homozygotes. Conjugation is a reciprocal fertilization producing genetically identical F1 heterozygotes (+/m) from each of the two parents. During autogamy of the F1s, only one of the parental alleles is retained and made homozygous, resulting in 50% of F2 clones of each genotype. (**C**) E revertants resulting from an E-inducing mutation will transmit the mutant allele to the F1 in the O cytoplasmic lineage, which remains O for recessive mutations. After autogamy of that F1, mutant F2 homozygotes will switch to E if the mutation is in a ZEG. (**D**) If the mutation is in a MEG, mutant F2 homozygotes remain O but give rise to E F3 progeny, after an additional autogamy. Shown here is the result expected for a maternal-effect mutation with a low penetrance, where only a fraction of F3 mutant homozygotes switch to E.

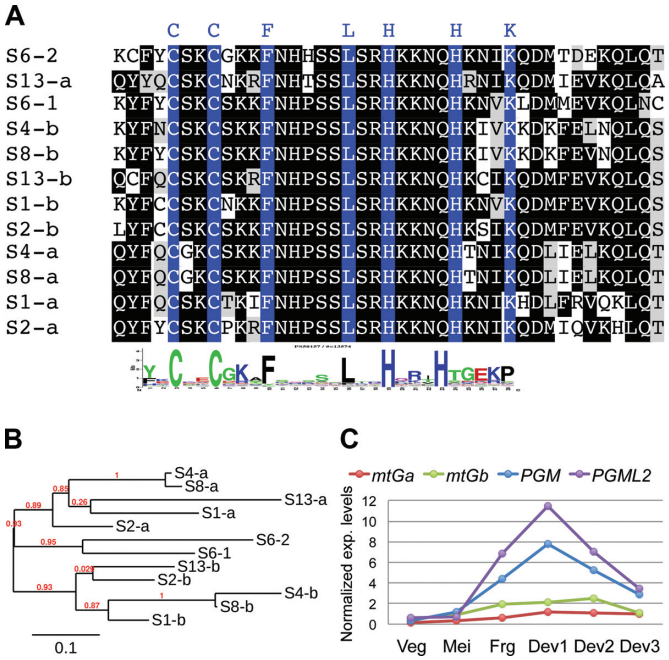

Figure 2. The *mtGa-1* mutation identifies a pair of ohnologous proteins with C2H2 zinc fingers. (**A**) Alignment of the mtG zinc fingers of the *a* and *b* ohnologs from 6 sibling species of the *P. aurelia* complex (S1, *P. primaurelia* strain Ir4-2; S2, *P. biaurelia* strain V1-4; S4, *P. tetraurelia* strain 51; S6, *P. sexaurelia* strain AZ8-4; S8, *P. octaurelia* strain 138; S13, *P. tredecaurelia*, strain 209). The logo of the Prosite motif PS50157 (C2H2_2) is shown below the alignment for comparison. (**B**) Phylogenetic analysis showing the clustering of *a* and *b* ohnologs, except for the early diverging *P. sexaurelia* where ohnologs are named *1* and *2*. (**C**) The expression profiles (normalized expression levels, arbitrary units) of *mtGa* and *mtGb* are compared to those of the *PGM* and *PGML2* genes during an autogamy time course (41): Veg, exponential vegetative cells; Mei, meiosis; Frg, fragmentation of the parental MAC; Dev1-Dev3, three stages of development of zygotic MACs.

E reversion at autogamy exceeds the frequency of mutants that can be recovered in these conditions.

Identification of mutations by whole-genome sequencing

The mutagenesis conditions used were previously estimated to result in an average of 30 mutations in the MAC-destined part of the genome of each viable clone, after a single autogamy of irradiated cells (28). To identify the putative E-inducing mutations and distinguish them from irrelevant, unlinked mutations, we resequenced the MAC genome of pools of F2 or F3 homozygotes with a mutant phenotype (see Methods). For the first putative mutant in a ZEG, a candidate mutation was identified in PTET.51.1.G0660190, a 426-bp gene encoding a protein with a C2H2 zinc finger (Figure 2A). The gene was named *mtGa*. The mutant allele, *mtGa-1*, is a null one, a frameshift mutation resulting from a 1-bp insertion in the fifth codon. *mtGa* has an ohnolog from the last WGD (PTET.51.1.G0710056, named *mtGb*) encoding a 65% identical, 80% similar protein. The two ohnologs are conserved in sibling species of the *P. aurelia* complex (Figure 2B) but not in the outgroups *P. multimicronucleatum* or *P. caudatum* (40). Microarray data (41) indicated that *mtGa* and *mtGb* are expressed specifically during MAC development, with expression profiles similar to that of the endonuclease Pgm (Figure 2C). For the second putative mutant in a MEG, a candidate mutation was found in the *PTIWI09* gene (PTET.51.1.G0660118), a C-to-T transition changing Serine 64 to Leucine. *PTIWI09* is indeed a MEG since it is only expressed during meiosis (12). The mutant allele was called *PTIWI09-1*.

Co-segregation of these mutations with the mating-type phenotypes was checked by genotyping series of individual F2 or F3 homozygotes in the above crosses (Supplementary Tables S3 and S4). The near-perfect correlations observed support the correct identification of the mutations. Nevertheless a few of the E F2s from the *mtGa-1* cross turned out to be wild-type homozygotes, indicating that these reversion events occurred in the absence of the mutation in the zygotic genome. This suggests that the reversion frequency may be higher after autogamy of *mtGa-1* heterozygotes than in the wild type. In the *PTIWI09-1* cross, genotyping revealed a small number of mutant homozygotes that had remained O in F3 populations, confirming the low penetrance of the mutation.

Genome-wide analysis of IES retention in the *PTIWI09-1* mutant

PCR tests confirmed that the *mtA* promoter was correctly excised in all individual F2 clones deriving from a mating type O *PTIWI09-1* heterozygote, including mutant homozygotes; after an additional autogamy, it was retained only in a fraction of individual F3 mutant homozygotes (5/14) (Supplementary Figure S1), explaining the selfer phenotype of F3 populations. The *PTIWI09-1* mutation has a weaker effect than *PTIWI09* silencing and is therefore a hypomorphic allele. The viability of the homozygous mutant rules out a strong effect on the excision of scnRNA-dependent IESs. In the mutant F3 MAC sequencing data the *mtA* promoter was retained in 42.7% of reads at one boundary and 41.6% at the other. Only 92 other IESs were found to be significantly more retained than in a wild-type MAC dataset, passing statistical tests based on the numbers of unexcised and excised reads (31). Furthermore, 85 of these did not pass these tests when compared to at least one of 6 other wild-type datasets. This suggests that most of these are cases of spontaneous retention (or IES annotation errors) rather than a true effect of the *PTIWI09-1* mutation.

The pipeline used for detection of IES retention is unable to assess the significance of low IES retention levels ($\leq$5% of the $\sim$800 genome copies in the MAC) even at high coverage (31), making it ill-adapted to document the effect of a hypomorphic allele. We therefore asked whether IESs with small, non-significant numbers of unexcised reads would be enriched in known small RNA-dependent IESs in the *PTIWI09-1* mutant, compared to wild-type datasets. For IESs showing only 1 or 2 unexcised reads at either boundary, no significant difference was observed, in line with the idea that occasional errors of the excision machinery (42) can be found for all classes of IESs. However, IESs with $\geq$3 unexcised reads showed a modest but significant enrichment in small RNA-dependent IESs in the *PTIWI09-1* data, while no difference was observed when two wild-type data sets were compared (Supplementary Figure S2).

PTIWI09-1 thus appears to have a very small effect on the efficiency of excision of small RNA-dependent IESs, barely detectable over the background of general excision errors.

mtGa and *mtGb* are both involved in excision of the *mtA* promoter

To confirm the requirement for *mtGa* and test for a possible role of *mtGb* in excision of the *mtA* promoter, both genes were silenced, separately or together, during autogamy of wild-type O cells. After autogamy, 30 cells were isolated from each silencing test, and the resulting post-autogamous clones were tested for mating types. In the 3 *mtG* silencing tests, all 90 clones had switched to E, while 12/12 clones from an unsilenced control remained O. A PCR test on DNA isolated from mass post-autogamous cells confirmed that the *mtA* promoter was retained in the new MACs (Supplementary Figure S3). Thus, both ohnologs are required for *mtA* promoter excision.

We further sought to complement the *mtGa-1* mutation with the wild-type gene. Backcrosses showed that once induced by the mutation, mating type E is self-maintained even in wild-type descendants, after the mutation is crossed out. Thus the only possible test is to provide the wild-type protein to mutant F2 homozygotes developing after autogamy of a mating type O *mtGa-1* heterozygote, and ask whether this will prevent the switch to E. However, it is not possible to microinject the developing MAC before the rearrangement stage, and consistent with genetic evidence that *mtGa* is mostly expressed from the zygotic genome, transformation of the maternal heterozygous MAC with an *mtGa*-GFP fusion construct did not result in any detectable fluorescence during autogamy, nor did it prevent mutant F2 progeny to switch to E (Supplementary Figure S4). We therefore replaced the endogenous *mtGa* promoter and 3′ UTR with those of the *PGML2* gene (Bischerour *et al.*, submitted), which has a similar expression profile but is expressed from the maternal MAC (Figure 2C). After autogamy of transformed mating type O heterozygotes, expression was verified by GFP fluorescence (see below), and all individual F2 clones (52/52) remained O. F2 clones, now devoid of the rescuing transgene, were then taken through an additional autogamy; as expected, about half of them gave rise to E F3 progeny, and genotyping confirmed that these were mutant homozygotes that had remained O in the F2 generation (Supplementary Figure S4). This demonstrates that the *mtGa-1* mutation is the cause of the mating type switch, and indicates that the mtGa-GFP fusion is functional in this regard.

mtGa colocalizes with H3K9me3 in the developing MAC

We took advantage of the above mtGa-GFP fusion to examine the subcellular localization of the protein throughout the different stages of autogamy. Only background fluorescence could be detected in vegetative cells and in the early stages of meiosis and fertilization, but a strong GFP signal accumulated soon after the second division of the zygotic nucleus in the two MAC anlagen. Initially appearing as a diffuse signal in the young anlagen (Figure 3A), GFP fluorescence then formed a punctate pattern of dozens

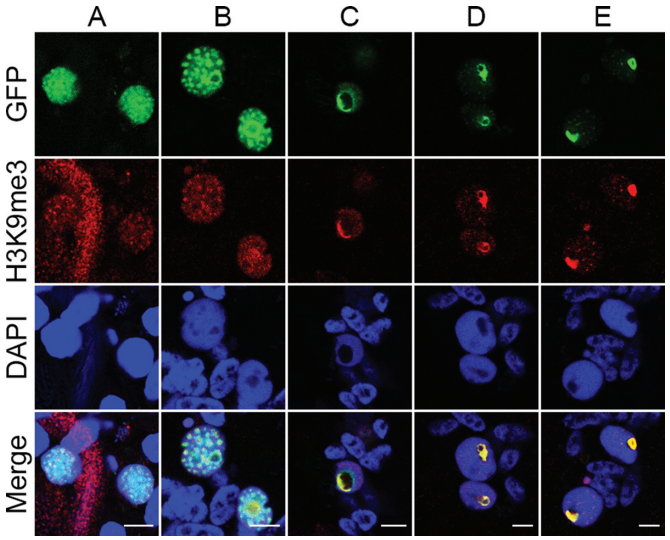

Figure 3. Localization of *mtGa*-GFP and H3K9me3 immunofluorescence at successive stages of MAC development. All pictures are projections of seven confocal optical sections through developing MACs, after autogamy of a clone transformed with pmtGa-GFP-L2. Representative images of successive developmental stages are shown from left to right (panels A–E). Autogamy cannot be induced simultaneously in all cells, and the developmental sequence was reconstructed by examining many cells at different time points. Fluorescence filters are indicated from top to bottom. Scale bars are 5 μm. (**A**) The GFP signal is initially diffuse in the nucleoplasm of developing MACs. The large trail seen in the H3K9me3 panel is due to unspecific labeling of the oral groove. The other nuclei stained by DAPI are fragments of the maternal MAC. (**B**) The signal progressively concentrates into a punctate pattern of vesicle-like structures with a dark center (see Supplementary Data S1) which appear to fuse together, until (**C**) a single large one is left. (**D**, **E**) The dark center shrinks in size, resulting in a single bright dot which finally disappears.

of vesicle-like foci (Figure 3B). Series of consecutive optical sections through the anlagen suggest that these foci are closed structures with dark centers that tend to concentrate at the nuclear periphery (Supplementary Data S1), similar to the DNA elimination structures that have been observed during MAC development in the related ciliate *Tetrahymena thermophila* (43–45). These foci later appeared to fuse, eventually forming a single large nuclear compartment delineated by GFP at its edge, with a dark center which did not stain with DAPI (Figure 3C). The size of the label-free inner region then decreased until the surrounding GFP signal formed a compact single dot (Figure 3D-E) which finally disappeared altogether.

These dynamic changes are also reminiscent of the localization of the histone methyltransferase protein Ezl1 and of H3K9me3 and H3K27me3 marks, which have been implicated in IES excision (18). Double labeling indeed showed that mtGa-GFP and H3K9me3 signals co-localized, most conspicuously during formation of the single large nuclear compartment (Figure 3C–E). Thus mtGa appears to associate with heterochromatin that is undergoing elimination.

mtGa and *mtGb* have largely redundant functions in excision of a small IES subset

The possible involvement of mtGa and mtGb in other DNA rearrangements was examined by resequencing the MAC

genomes of mass post-autogamous progeny from wild-type mating type O cultures depleted for these two proteins by RNAi, either separately or together (Supplementary Table S5). RNAi efficiency was checked by testing 30 individual clones from each progeny, which were all mating type E. IES retention was assessed genome-wide with a pipeline comparing the numbers of processed and unprocessed reads at each of the two boundaries independently (31) (see Methods). Because few IESs were affected, we used stringent criteria to avoid including spurious cases of spontaneous retention or IES annotation errors: we counted only IESs that showed significantly higher retention levels when compared to each of seven wild-type data sets. In addition to the *mtA* promoter, 105 IESs were found to be retained in the double knockdown, most often with a high retention score (0.70 on average) (Supplementary Table S6). In 14 cases, only one of the two boundaries was strongly affected while the other was cleaved and rejoined with an alternative TA inside the IES, resulting in MAC retention of only part of the IES.

In the single mtGa knockdown, only 40 IESs were retained (all among the 105), while the single mtGb knockdown impaired excision of only 14 of these and an additional one. Analysis of the MAC genome of *mtGa-1* mutant pools was fairly consistent: this revealed 31 retained IESs, 26 of which were retained in the single mtGa knockdown. However, most were significantly retained only in the F3 generation and did not pass the statistical tests in the F2 data, suggesting that maternal contribution to *mtGa* expression (from the F1 heterozygous MAC) is usually sufficient to compensate for the lack of zygotic expression of the *mtGa* ohnolog. Only four IESs behaved like the *mtA* promoter and were significantly retained in the F2 generation (Supplementary Table S6). We conclude that mtGa and mtGb are largely redundant in their excision function, though some of the 105 IESs specifically require mtGa or even both ohnologs, as is the case of the *mtA* promoter.

mtG-dependent IESs have large sizes and specific end sequences

The surprising finding that a pair of proteins is required for excision of such a small number of IESs (~0.2% of the total) prompted us to search for specific features that might distinguish that subset. mtG-dependent IESs are on the high end of the size distribution, ranging from 85 to 500 bp with a median size of 214 bp, versus 50 bp for the general set (2). They do not differ from other IESs of similar sizes in G+C content (19% overall) or distribution of insertion sites (68% exons, 12% introns, 4% 5′UTRs, 2% 3′UTRs, 14% intergenic regions). The 90 genes interrupted by mtG-dependent IESs, however, may not be random: 30 of them encode proteins with transmembrane helices (33% versus 17% of all protein-coding genes).

We examined the genetic requirements for excision of mtG-dependent IESs by looking at their retention levels in sequencing data sets from published knockdowns of different factors involved in small RNA pathways or histone modifications. mtG-dependent IESs were more often retained to significant levels than genome-wide IESs (Figure 4A and Supplementary Data S2). Since the fraction of IESs affected in these knockdowns is known to increase with IES

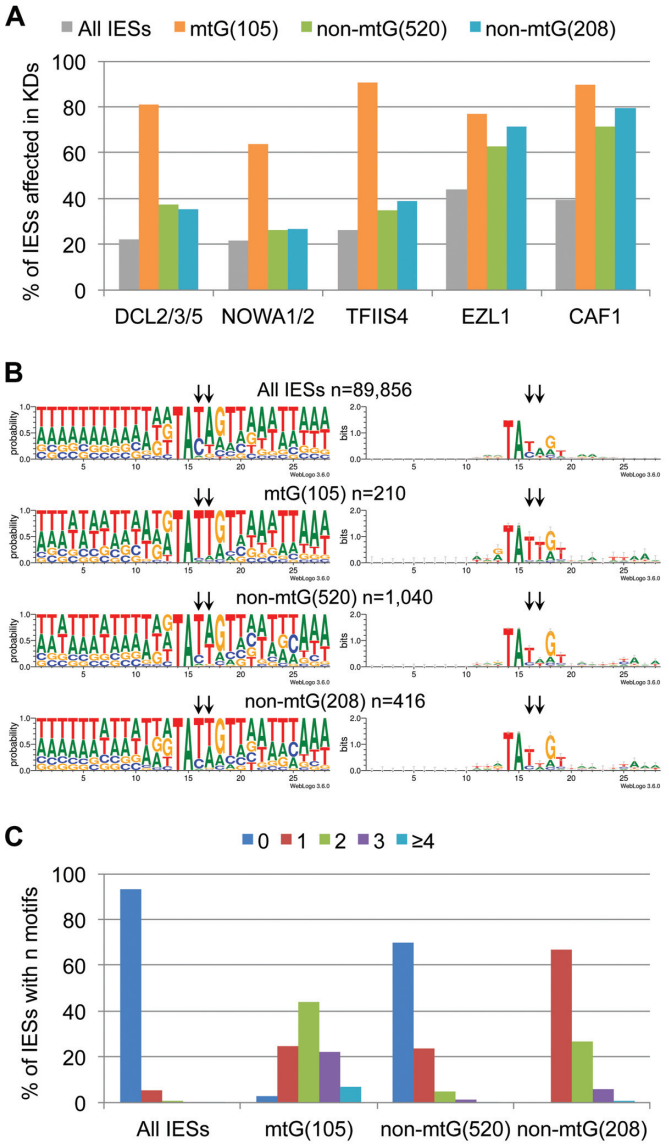

Figure 4. Characteristics of *mtG*-dependent IESs. (**A**) Fraction of IESs (%) that are sensitive to the knockdown (KD) of different factors involved in IES excision in different subsets: mtG(105), the 105 IESs retained in the double *mtG* knockdown; non-mtG(520), a size-matched control set of 520 *mtG*-independent IESs; non-mtG(208), a size-matched control set of 208 *mtG*-independent IESs containing at least one GCTAA motif each. (**B**) Logos of the 13 bases on either side of the TA boundaries (left and right combined) of the different IES sets. Arrows point to the first two positions inside IESs, where base composition is significantly different in *mtG*-dependent IESs. (**C**) Fractions of IESs (%) with n GCTAA motifs ($n = 0$, 1, 2, 3, ≥4) in the different IES sets.

size (11,17–19), we looked at a control set of 520 IESs with the same size distribution, but completely excised in all of the mtG silencing or mutant data sets. Compared to these, mtG-dependent IESs were still more than twice as likely to be affected in knockdowns of the small RNA-pathway genes *DCL2/3/5*, *NOWA1/2*, and *TFIIS4* (81%, 64% and 90%, respectively). The same trend was observed after silencing of *DCL2/3*, but in contrast to the *mtA* promoter, none of the mtG-dependent IESs was significantly retained

in the single *DCL5* knockdown (not shown). In contrast, the high retention frequencies in knockdowns of genes involved in chromatin modifications (*EZL1*, *CAF1*) are also seen in the control set and are therefore attributable to large IES size.

The end sequences of mtG-dependent IESs show a striking deviation from the general consensus. Instead of the globally more frequent 5′-TAYAR-3′, 74% of the 210 individual ends have the sequence 5′-TATTR-3′ (Figure 4B), as do both ends of the *mtA* promoter segment excised in mating type O clones. The frequency of that end sequence (13% in all IESs) is known to be higher among long IESs (>150 bp) (46), but in the size-matched control it reached only 34% of IES ends. The most discriminating feature appeared to be the presence of a pyrimidine at the second position 3′ of the TA (TANY): this occurred in 90% of IES ends in the mtG-dependent set (versus 40% in the size-matched control), and only one of the 105 IESs had a purine at both ends. Furthermore, within the mtG-dependent set individual retention scores in the mtGa+b knockdown were higher for TANY ends (average 0.72) than for TANR ends (average 0.55).

mtG-dependent IESs are enriched in a 5-bp motif at internal positions

We then examined the internal sequences of mtG-dependent IESs. The DREME tool (37) revealed a significant over-representation of the 5-bp motif GCTAA/TTAGC relative to the size-matched, mtG-independent control set. On average, mtG-dependent IESs contain 2.10 such motifs per IES (overall motif density 9.28/kb) (Figure 4C), and only 3 of the 105 IESs lack any occurrence. The *mtA* promoter contains two. In the size-matched control set, 70% of the 520 IESs lack the motif (0.39 motif per IES on average), and overall motif density is 1.72/kb. In the whole set of 44 928 IESs, only 7% contain the motif (0.10 motif per IES on average); overall motif density is 1.30/kb, close to the value expected by chance given IES base composition (1.33/kb). Outside of IESs, this motif is slightly under-represented in the entire MAC genome, and a little more so in intergenic regions (92 and 85% of the values expected from base composition, respectively). A DREME analysis did not identify any significantly enriched or depleted motif in the 100 bp of MAC sequences flanking mtG-dependent IESs on either side, relative to those of the size-matched set.

The motif also shows some association with TATTR boundaries in mtG-independent IESs. In a second control set of 208 IESs with the same size distribution, showing no retention in any of the mtG silencing or mutant data sets but selected to contain at least one motif each (overall motif density 6.14/kb), the frequencies of TATTR and TANY ends reached 49% and 55%, respectively. Reciprocally, IESs in the size range of mtG-dependent IESs (85–500 bp) with TATTR (or TANY) at both ends have an overall motif density of 2.65/kb (or 2.63/kb), versus 1.47/kb in all 85–500-bp IESs. Thus this association of features is not sufficient to predict mtG-dependent IESs with good specificity. For instance, the subset defined as 'IESs between 85 and 500 bp in size, with at least one TANY end and at least one motif' does include 101 of the 105 mtG-dependent IESs, but the total number of such IESs in the genome is 860. This suggests that some additional, unidentified IES property determines mtG requirement for excision.

Within mtG-dependent IESs, the GCTAA motif is not randomly distributed. Its frequency remains relatively low over the first 43 bp internal to both TA ends, close to the frequency observed in the first size-matched control set, and then rises sharply on both strands starting from position 44 (Figure 5A). Interestingly, no abrupt change in density is seen in the mtG-independent size-matched sets, or in the 759 motif-containing, 85–500-bp, mtG-independent IESs with at least one TANY end, indicating that the motif is homogeneously distributed up to the ends in these sets. While this observation suggests a position-dependent role for the motif in mtG-dependent excision, it probably does not help much to improve prediction specificity, since many mtG-independent IESs do contain motifs in their central portions.

The C2H2 zinc fingers of mtGa and mtGb have identical amino acids at the positions thought to participate in DNA binding. For both proteins, the best binding sites are predicted to be GCTA or GCTT by the 'polynomial' or 'linear expanded' support vector machines, respectively, described in (38) (Figure 5B and Supplementary Figure S6). This strongly supports the idea that these proteins promote IES excision through direct binding to the GCTAA motif, and further suggests that they may also bind some degenerate forms. GCTAT, GCTAG, and GCTAC are also over-represented in mtG-dependent IESs, though at more moderate levels (155–280% of expected frequencies). Globally, the 295 occurrences of GCTA represent 429% of the expected number, and the same abrupt increase in density is seen on both strands after the first 44 bp in mtG-dependent IESs, but not in control sets (Supplementary Figure S7A). While GCTT occurrences are much less frequent (121% of expected), local density follows the same trend, raising the possibility that it also plays a role in mtG-dependent excision (Supplementary Figure S7B and Figure 5C for GCTW). Binding to variant sites might thus explain how mtG proteins are involved in the excision of three IESs that lack any GCTAA motif.

Evolutionary conservation of mtGa and mtGb function

25 of the 105 mtG-dependent IESs are inserted into genes that have retained their duplicate from the last WGD, and have 'ohnologous' IESs inserted at the same positions in the duplicates (2). In only one of these pairs are both ohnologs mtG-dependent. That case is also unusual in that ohnologous IES sequences have retained 82% sequence identity, while most IES pairs from the last WGD have lost all sequence similarity (2). In that case, the alignment shows that a single GCTAA site is conserved in both IES ohnologs. In the 23 discordant pairs, the mtG-independent ohnolog contains a motif in only seven cases. In three of them the motifs are on opposite strands in the mtG-dependent and -independent ohnologs, making it likely that these represent independent gains or losses of motifs since the last

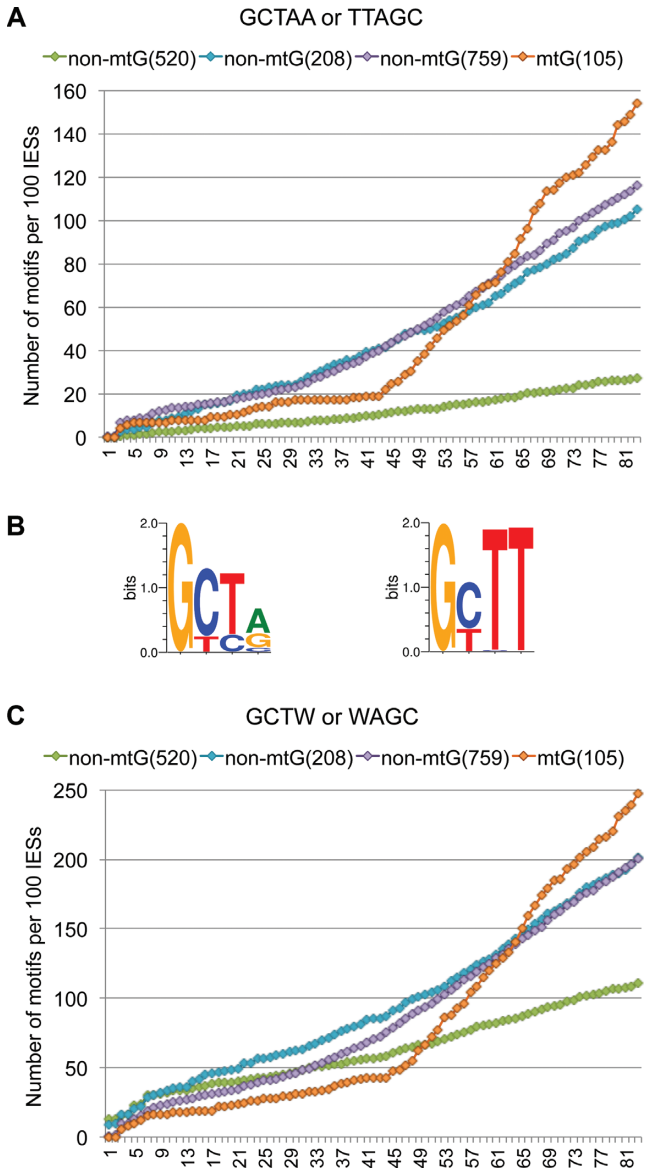

Figure 5. Distribution of motifs within *mtG*-dependent IESs. (**A**) Cumulative distribution of the numbers of GCTAA or TTAGC motifs over the first 83 positions internal to the TA boundaries (normalized, per 100 IESs in each set). Both left and right ends are analyzed together. The abrupt change in density starting at position 44 in *mtG*-dependent IESs (mtG(105)) is not seen in the *mtG*-independent control sets, including those selected to contain at least one motif in each IES (non-mtG(208) and non-mtG(759)). The same change is seen when GCTAA and TTAGC are analyzed separately (Supplementary Figure S5). (**B**) Prediction of best binding sites for the C2H2 zinc finger of mtG proteins by the polynomial (left) and linear expanded (right) support vector machines described in (38) (see Supplementary Figure S6 for base-probability matrices). (**C**) Cumulative distribution of the numbers of GCTW or WAGC motifs, which appear to be under-represented in the first 43 positions of *mtG*-dependent IESs.

WGD, rather than conservation from the ancestral IES sequence. Since IES end sequences are well conserved between ohnologs, this analysis appears to confirm that the presence of a GCTAA motif, while apparently necessary, is not sufficient to determine mtG requirement.

We then examined the orthologs of a few IESs in the closely related species *P. octaurelia* (strain 138), which diverged from *P. tetraurelia* long after the last WGD in their common ancestor. Sequence identity (58–88%) was sufficient to assess conservation of the *P. tetraurelia* motifs, and ask whether this correlates with conservation of mtG requirement. The *mtA* promoter differs by only two substitutions in *P. octaurelia* and both GCTAA motifs are conserved (25); silencing of *mtGa* and *mtGb* during autogamy of mating type O cells indeed caused progeny to retain the *mtA* promoter and to switch to E, as in *P. tetraurelia* (Supplementary Figure S8A). Retention levels were also tested for the orthologs of eight IESs that are mtG-dependent in *P. tetraurelia* (Supplementary Figure S8B and Supplementary Table S7). In three cases (IESs 1–3), the *P. octaurelia* orthologous IESs were partially retained after depletion of mtGa and/or mtGb, all showing some incorrect excision using alternative TA boundaries internal to the IES sequences. In these IESs four of six of the *P. tetraurelia* motifs were conserved, and there were three additional occurrences (Supplementary Figure S9). However, in three other cases (IESs 4–6) the *P. octaurelia* orthologs were correctly excised, despite conservation of five of the six *P. tetraurelia* motifs. In the last two cases (IESs 7–8), both annotated as intergenic IESs in *P. tetraurelia*, the orthologous sequences in *P. octaurelia* are not excised in the MACs of wild-type cells and are therefore not IESs; both lack the motif. Thus, although the GCTAA motif appears to be necessary for mtG-dependent excision, this analysis confirms that it is not sufficient. mtG requirement is further shown to be evolutionarily labile, having changed in half of IESs since the divergence of these sibling species.

DISCUSSION

In this work we have tested the feasibility of a mutagenesis screen to identify new genes involved in the scnRNA pathway and/or IES excision machinery in *P. tetraurelia*. The screen is based on the fact that excision of the *mtA* promoter, like maternally inherited deletions of other cellular genes or gene parts, is more strongly dependent on the scnRNA pathway than is excision of most IESs (10,12); many IESs, and in particular the oldest ones, appear to have acquired sequence features and/or epigenetic marks that promote excision in a scnRNA-independent manner. The two mutations recovered provide a proof-of-principle that the screen can identify essential genes through non-lethal hypomorphic alleles, or genes with redundant WGD ohnologs, which make up half of all genes but are elusive mutagenesis targets. One was found to be a missense mutation in the *PTIWI09* gene, which together with its ohnolog *PTIWI01* is essential for correct genome rearrangements and viability (12,13). The other identified a new pair of ohnologous genes, *mtGa/b*, which suggest previously unsuspected roles for sequence-specific DNA-binding proteins in IES recognition and/or excision.

The productivity of the screen, however, was limited by two factors. First, with only two mutants identified out of 60 mating type revertants tested, the rate of spontaneous mating type change after autogamy turned out to be too high in the cell line and conditions used, making the screening te-

dious and time-consuming. Secondly, the need for two successive rounds of autogamy to reveal mutations in MEGs likely results in the loss of many mutants – and certainly of slow-growing ones – because it is impossible to expand the entire progeny of the first autogamy for the ∼20 divisions required to take them through the immaturity period. Improving the yield would thus depend on the identification of a cell line in which spontaneous mating type change is much reduced and autogamy can be induced at a very young clonal age, or of experimental treatments resulting in the same effects.

The *PTIWI09-1* mutation changes a serine (Ser64Leu) that is highly conserved in the N domain of *Paramecium* Piwi proteins (12). Often a threonine in eukaryotic Argonaute proteins, this residue is next to a conserved tyrosine which was shown in human Ago2 (Tyr101) to be involved in the unwinding of the passenger strand, after loading of small RNA duplexes (47). Interestingly, this Ser/Thr residue is naturally a leucine in many metazoan Piwi proteins, which may be related to the fact that piRNAs, unlike ciliate scnRNAs, are not produced by Dicer-mediated cleavage of dsRNA precursors. The *PTIWI09-1* allele has a very subtle phenotypic effect, weaker than the single silencing of the *PTIWI09* gene, suggesting that the mutant protein is partially functional. The increased frequency of O-to-E mating type change, and occasional retention of scnRNA-dependent IESs, might be attributable to slightly impaired unwinding of scnRNA passenger strands. Whatever the exact molecular defect may be, the recovery of a viable mutant in such a central actor of the scnRNA pathway suggests that additional factors may be uncovered through mutagenesis.

The second mutation identified, *mtGa-1*, allowed us to define for the first time a mechanistically distinct subset of IESs characterized by a specific DNA motif (GCTAA) at internal positions. In *T. thermophila*, the Lia3 protein was shown to be required for correct excision of a subset of IESs (48), although in that case the protein appears to bind G-quadruplex structures formed by polypurine tracts that are located outside of IESs, thus defining proper excision boundaries. The C2H2 zinc finger present in both mtGa and mtGb is predicted to bind GCTA or GCTT, suggesting that direct binding to IESs is required for correct excision. We currently cannot exclude the possibility that mtG proteins bind to the TFIIS4-dependent nascent transcripts found to play a role in IES excision (15), rather than to DNA itself. In both cases, the co-localization of mtGa-GFP and H3K9me3 marks in what appears to be DNA elimination structures would suggest that binding is maintained until excision. This first example raises the possibility that other subsets of genome-wide IESs may similarly depend on the binding of different protein factors to different motifs, explaining why functionally relevant motifs have not been detected so far.

The precise function of mtG proteins remains unclear. Approximately 80% of the 105 mtG-dependent IESs also depend on Dcl2/3/5, TFIIS4, and Nowa1/2 for excision, indicating that mtG function does not substitute for small RNA-mediated recognition. Thus, unlike the KRAB zinc-finger proteins shown to target heterochromatin formation on TEs in mammals (49), mtG binding does not appear to be generally required for direct recruitment of Ezl1 activity. Instead it could be required in some cases, depending on a variable chromatin property such as nucleosome positioning, for assembly of the excision complex over the correct TA boundaries. This is suggested by cases where mtG depletion results in activation of cryptic TA boundaries rather than retention of the entire IES, and by indirect evidence for a position-dependent role of GCTAA motifs.

The non-random sizes, end sequences, and motif density of mtG-dependent IESs are not sufficient to predict mtG requirement, since many similar IESs are correctly excised in mtG knockdowns. While this may still depend on unidentified sequence features such as intrinsic DNA curvature, presence or absence of cryptic TA boundaries, or binding sites for other, functionally redundant factors, an intriguing possibility is that mtG binding depends on the methylation state of the GCTAA motif, as reported for zinc-finger proteins in other systems (50,51). 5-methylcytosine has not been detected in *P. tetraurelia*, but N6-methyladenine is abundant (52). Our results indicate that individual IESs have frequently gained or lost mtG requirement during evolution, including in cases where the motifs are conserved (but may have changed methylation status). The system is unlikely to have been conserved in *P. aurelia* species just for the sake of excision of a handful of IESs, unless it serves some regulatory function. mtG activity is not essential in laboratory conditions and might be naturally regulated by environmental conditions; this could be used to coordinately inactivate specific sets of genes, defined in a flexible manner by the methylation state of individual motifs. Whatever their roles may be, the implication of sequence-specific factors in excision of a small subset of IESs provides a new paradigm for the problem of IES recognition, and suggests we may just be seeing the tip of the iceberg.

DATA AVAILABILITY

All datasets generated in this study were deposited in the European Nucleotide Archive under the Project Accession PRJEB27011. The accession number of each dataset can be found in Supplementary Table S5. Gene accession numbers refer to the latest annotation of the *P. tetraurelia* strain 51 genome (53).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank A. Potekhin for providing *P. octaurelia* strain 138, G. Pellerin for help with microinjections, V. Tanty and N. Stahlberger for technical support. We thank S. Marker and S. Malinsky and all lab members for continuous support and discussions. The sequencing benefited from the facilities and expertise of the high-throughput sequencing platform of I2BC. We acknowledge the IBENS Imaging Facility, member of the national infrastructure France-BioImaging [ANR-10-INBS-04], where confocal imaging was carried out.

FUNDING

Agence Nationale de la Recherche [ANR-12-BSV6-0017-04 INFERNO to E.M. and L.S.]; Fondation pour la Recherche Médicale [Equipe FRM DEQ20150331763 to E.M.]; European Research Council (ERC) [260358 'EPIGENOME' and 681178 'G-EDIT' to M.N.]; Swiss National Science Foundation [31003A_146257 and 31003A_166407 to M.N]; 'Investissements d'Avenir' launched by the French Government and implemented by ANR with the references [ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL]. Funding for open access charge: Centre National de la Recherche Scientifique.

Conflict of interest statement. None declared.

REFERENCES

1. Betermier,M. and Duharcourt,S. (2014) Programmed rearrangement in ciliates: Paramecium. *Microbiol.Spectrum*, **2**, MDNA3-0035-2014.
2. Arnaiz,O., Mathy,N., Baudry,C., Malinsky,S., Aury,J.M., Wilkes,C.D., Garnier,O., Labadie,K., Lauderdale,B.E., Le Mouel,A. *et al.* (2012) The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLos Genet.*, **8**, e1002984.
3. Dubois,E., Bischerour,J., Marmignon,A., Mathy,N., Regnier,V. and Betermier,M. (2012) Transposon invasion of the paramecium germline genome countered by a domesticated PiggyBac transposase and the NHEJ pathway. *Int. J. Evol. Biol.*, **2012**, 436196.
4. Baudry,C., Malinsky,S., Restituito,M., Kapusta,A., Rosa,S., Meyer,E. and Betermier,M. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev.*, **23**, 2478–2483.
5. Dubois,E., Mathy,N., Regnier,V., Bischerour,J., Baudry,C., Trouslard,R. and Betermier,M. (2017) Multimerization properties of PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements. *Nucleic Acids Res.*, **45**, 3204–3216.
6. Kapusta,A., Matsuda,A., Marmignon,A., Ku,M., Silve,A., Meyer,E., Forney,J.D., Malinsky,S. and Betermier,M. (2011) Highly precise and developmentally programmed genome assembly in Paramecium requires ligase IV-dependent end joining. *PLos Genet.*, **7**, e1002049.
7. Marmignon,A., Bischerour,J., Silve,A., Fojcik,C., Dubois,E., Arnaiz,O., Kapusta,A., Malinsky,S. and Betermier,M. (2014) Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate Paramecium tetraurelia. *PLos Genet.*, **10**, e1004552.
8. Klobutcher,L.A. and Herrick,G. (1995) Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. *Nucleic Acids Res.*, **23**, 2006–2013.
9. Hoehener,C., Hug,I. and Nowacki,M. (2018) Dicer-like enzymes with sequence cleavage preferences. *Cell*, **173**, 234–247.
10. Lepere,G., Nowacki,M., Serrano,V., Gout,J.F., Guglielmi,G., Duharcourt,S. and Meyer,E. (2009) Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. *Nucleic Acids Res.*, **37**, 903–915.
11. Sandoval,P.Y., Swart,E.C., Arambasic,M. and Nowacki,M. (2014) Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. *Dev. Cell*, **28**, 174–188.
12. Bouhouche,K., Gout,J.F., Kapusta,A., Betermier,M. and Meyer,E. (2011) Functional specialization of Piwi proteins in Paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling. *Nucleic Acids Res.*, **39**, 4249–4264.
13. Furrer,D.I., Swart,E.C., Kraft,M.F., Sandoval,P.Y. and Nowacki,M. (2017) Two sets of piwi proteins are involved in distinct sRNA pathways leading to elimination of Germline-Specific DNA. *Cell Rep.*, **20**, 505–520.
14. Lepere,G., Betermier,M., Meyer,E. and Duharcourt,S. (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. *Genes Dev.*, **22**, 1501–1512.
15. Maliszewska-Olejniczak,K., Gruchota,J., Gromadka,R., Denby Wilkes,C., Arnaiz,O., Mathy,N., Duharcourt,S., Betermier,M. and Nowak,J.K. (2015) TFIIS-Dependent Non-coding transcription regulates developmental genome rearrangements. *PLos Genet.*, **11**, e1005383.
16. Nowacki,M., Zagorski-Ostoja,W. and Meyer,E. (2005) Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. *Curr. Biol.: CB*, **15**, 1616–1628.
17. Swart,E.C., Denby Wilkes,C., Sandoval,P.Y., Hoehener,C., Singh,A., Furrer,D.I., Arambasic,M., Ignarski,M. and Nowacki,M. (2017) Identification and analysis of functional associations among natural eukaryotic genome editing components [version 1; referees: 1 approved, 1 approved with reservations]. *F1000Research*, **6**, 1374.
18. Lhuillier-Akakpo,M., Frapporti,A., Denby Wilkes,C., Matelot,M., Vervoort,M., Sperling,L. and Duharcourt,S. (2014) Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLos Genet.*, **10**, e1004665.
19. Ignarski,M., Singh,A., Swart,E.C., Arambasic,M., Sandoval,P.Y. and Nowacki,M. (2014) Paramecium tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination. *Nucleic Acids Res.*, **42**, 11952–11964.
20. Duharcourt,S., Butler,A. and Meyer,E. (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on Paramecium tetraurelia. *Genes Dev.*, **9**, 2065–2077.
21. Duharcourt,S., Keller,A.M. and Meyer,E. (1998) Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in Paramecium tetraurelia. *Mol. Cell. Biol.*, **18**, 7075–7085.
22. Allen,S.E., Hug,I., Pabian,S., Rzeszutek,I., Hoehener,C. and Nowacki,M. (2017) Circular concatemers of Ultra-Short DNA segments produce regulatory RNAs. *Cell*, **168**, 990–999.
23. Brygoo,Y. and Keller,A.M. (1981) Genetic analysis of mating type differentiation in Paramecium tetraurelia. III. A mutation restricted to mating type E and affecting the determination of mating type. *Dev. Genet.*, **2**, 13–22.
24. Meyer,E. and Keller,A.M. (1996) A Mendelian mutation affecting mating-type determination also affects developmental genomic rearrangements in Paramecium tetraurelia. *Genetics*, **143**, 191–202.
25. Singh,D.P., Saudemont,B., Guglielmi,G., Arnaiz,O., Gout,J.F., Prajer,M., Potekhin,A., Przybos,E., Aubusson-Fleury,A., Bhullar,S. *et al.* (2014) Genome-defence small RNAs exapted for epigenetic mating-type inheritance. *Nature*, **509**, 447–452.
26. Beisson,J., Betermier,M., Bre,M.H., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. Jr *et al.* (2010) Mass culture of Paramecium tetraurelia. *Cold Spring Harbor Protoc.*, **2010**, pdb.prot5362.
27. Beisson,J., Betermier,M., Bre,M.H., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. Jr *et al.* (2010) Maintaining clonal Paramecium tetraurelia cell lines of controlled age through daily reisolation. *Cold Spring Harbor Protoc.*, **2010**, pdb.prot5361.
28. Marker,S., Carradec,Q., Tanty,V., Arnaiz,O. and Meyer,E. (2014) A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia. *Nucleic Acids Res.*, **42**, 7268–7280.
29. Skouri,F. and Cohen,J. (1997) Genetic approach to regulated exocytosis using functional complementation in Paramecium: identification of the ND7 gene required for membrane fusion. *Mol. Biol. Cell*, **8**, 1063–1071.
30. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
31. Denby Wilkes,C., Arnaiz,O. and Sperling,L. (2016) ParTIES: a toolbox for Paramecium interspersed DNA elimination studies. *Bioinformatics*, **32**, 599–601.
32. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery Rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
33. Beisson,J., Betermier,M., Bre,M.H., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. Jr *et al.* (2010) Silencing specific Paramecium tetraurelia genes by feeding

double-stranded RNA. *Cold Spring Harbor Protoc.*, **2010**, pdb.prot5363.
34. Beisson,J., Betermier,M., Bre,M.H., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. Jr *et al.* (2010) DNA microinjection into the macronucleus of paramecium. *Cold Spring Harbor Protoc.*, **2010**, pdb.prot5364.
35. Beisson,J., Betermier,M., Bre,M.H., Cohen,J., Duharcourt,S., Duret,L., Kung,C., Malinsky,S., Meyer,E., Preer,J.R. Jr *et al.* (2010) Immunocytochemistry of Paramecium cytoskeletal structures. *Cold Spring Harbor Protoc.*, **2010**, pdb.prot5365.
36. Dereeper,A., Guignon,V., Blanc,G., Audic,S., Buffet,S., Chevenet,F., Dufayard,J.F., Guindon,S., Lefort,V., Lescot,M. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.
37. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
38. Persikov,A.V. and Singh,M. (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.
39. Brygoo,Y. (1977) Genetic analysis of mating-type differentiation in Paramecium tetraurelia. *Genetics*, **87**, 633–653.
40. McGrath,C.L., Gout,J.F., Doak,T.G., Yanagi,A. and Lynch,M. (2014) Insights into three whole-genome duplications gleaned from the Paramecium caudatum genome sequence. *Genetics*, **197**, 1417–1428.
41. Arnaiz,O., Gout,J.F., Betermier,M., Bouhouche,K., Cohen,J., Duret,L., Kapusta,A., Meyer,E. and Sperling,L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. *BMC Genomics*, **11**, 547.
42. Duret,L., Cohen,J., Jubin,C., Dessen,P., Gout,J.F., Mousset,S., Aury,J.M., Jaillon,O., Noel,B., Arnaiz,O. *et al.* (2008) Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline. *Genome Res.*, **18**, 585–596.
43. Madireddi,M.T., Coyne,R.S., Smothers,J.F., Mickey,K.M., Yao,M.C. and Allis,C.D. (1996) Pdd1p, a novel chromodomain-containing protein, links heterochromatin assembly and DNA elimination in Tetrahymena. *Cell*, **87**, 75–84.
44. Madireddi,M.T., Davis,M.C. and Allis,C.D. (1994) Identification of a novel polypeptide involved in the formation of DNA-containing vesicles during macronuclear development in Tetrahymena. *Dev. Biol.*, **165**, 418–431.
45. Nikiforov,M.A., Gorovsky,M.A. and Allis,C.D. (2000) A novel chromodomain protein, pdd3p, associates with internal eliminated sequences during macronuclear development in Tetrahymena thermophila. *Mol. Cell. Biol.*, **20**, 4128–4134.
46. Swart,E.C., Wilkes,C.D., Sandoval,P.Y., Arambasic,M., Sperling,L. and Nowacki,M. (2014) Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Res.*, **42**, 8970–8983.
47. Kwak,P.B. and Tomari,Y. (2012) The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat. Struct. Mol. Biol.*, **19**, 145–151.
48. Carle,C.M., Zaher,H.S. and Chalker,D.L. (2016) A parallel G Quadruplex-Binding protein regulates the boundaries of DNA elimination events of tetrahymena thermophila. *PLos Genet.*, **12**, e1005842.
49. Yang,P., Wang,Y. and Macfarlan,T.S. (2017) The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet.*, **33**, 871–881.
50. Blattler,A. and Farnham,P.J. (2013) Cross-talk between site-specific transcription factors and DNA methylation states. *J. Biol. Chem.*, **288**, 34287–34294.
51. Jin,J., Lian,T., Gu,C., Yu,K., Gao,Y.Q. and Su,X.D. (2016) The effects of cytosine methylation on general transcription factors. *Sci. Rep.*, **6**, 29119.
52. Cummings,D.J., Tait,A. and Goddard,J.M. (1974) Methylated bases in DNA from Paramecium aurelia. *Biochim. Biophys. Acta*, **374**, 1–11.
53. Arnaiz,O., Van Dijk,E., Betermier,M., Lhuillier-Akakpo,M., de Vanssay,A., Duharcourt,S., Sallet,E., Gouzy,J. and Sperling,L. (2017) Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression. *BMC Genomics*, **18**, 483.